Multi-modal Analysis of Music: A large-scale Evaluation

Rudolf Mayer
Institute of Software Technology and Interactive Systems
Vienna University of Technology
Vienna, Austria
mayer@ifs.tuwien.ac.at

Robert Neumayer
Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
neumayer@idi.ntnu.no

***Abstract*—Multimedia data by definition comprises several different types of content modalities. Music specifically inherits e.g. *audio* at its core, *text* in the form of lyrics, *images* by means of album covers, or *video* in the form of music videos. Yet, in many Music Information Retrieval applications, only the audio content is utilised. Recent studies have shown the usefulness of incorporating other modalities; in most of them, textual information in the form of song lyrics or artist biographies, were employed. Following this direction, the contribution of this paper is a large-scale evaluation of the combination of audio and text (lyrics) features for genre classification, on a database comprising over 20,000 songs. We present the audio and lyrics features employed, and provide an in-depth discussion of the experimental results.**

I. Introduction

With the ever-growing spread of music available in digital formats – be it in online music stores or on consumers' computer or mobile music players – Music Information Retrieval (MIR) as a research area dealing with ways to organise, structure and retrieve such music, is of increasing importance. Many of its typical tasks such as genre classification or similarity retrieval / recommendation often rely on only one of the many modalities of music, namely the audio content itself. However, music comprises many more different modalities. Text is present in the form of song lyrics, as well as artist biographies or album reviews, etc. Many artists and publishers put emphasis on carefully designing an album cover to transmit a message coherent with the music it represents. Similar arguments also hold true for music videos.

Recent research has to some extent acknowledged the multi-modality of music, with most research studies focusing on lyrics for e.g. emotion, mood or topic detection. In this paper, we apply our previous work on extracting rhyme and style features from song lyrics, with the goal of improving genre classification. Our main contribution is a large-scale evaluation on a database comprising over 20.000 songs from various different genres. Our goal in this paper is to show the applicability of our techniques to, and the potential of lyrics-based features on a larger test collection.

The remainder of this paper is structured as follows. In Section II, we briefly review related work in the field of multi-modal music information retrieval. Section III will outline the audio and lyrics features employed in our study. In Section IV, we describe our test collection and its most interesting properties, while in Section V we discuss the results in genre classification on this collection. Finally, Section VI will give conclusions and an outlook on future work.

II. Related Work

Music Information Retrieval is a sub-area of information retrieval, concerned with adequately organising, structuring and accessing (digital) audio. Important research directions include for example similarity retrieval, musical genre classification, or music analysis and knowledge representation. Comprehensive overviews of the research field are given in [1], [2].

The still dominant method of processing audio files in music information retrieval is by analysis of the audio signal, which is computed from plain wave files or via a preceding decoding step from other wide-spread audio formats such as MP3 or the (lossless) Flac format. A wealth of different descriptive features for the abstract representation of audio content have been presented. Early overviews on content-based music information retrieval and experiments is given in [3] and [4], focussing mainly on automatic genre classification of music. In this work, we employ mainly the Rhythm Patterns, Rhythm Histograms and Statistical Spectrum Descriptors [5], which we will discuss in more detail in Section III. Other feature sets include e. g. MPEG-7 audio descriptors or MARSYAS.

Several research teams have further begun working on adding textual information to the retrieval process, predominantly in the form of song lyrics and an abstract vector representation of the term information contained in text documents. A semantic and structural analysis of song lyrics is conducted in [6]. The definition of artist similarity via song lyrics is given in [7]. It is pointed out that acoustic is superior to textual similarity, yet that a combination of both approaches might lead to better results. Another area were lyrics have been employed is the field of emotion detection and classification, for example [8], which aims at disambiguating music emotion with lyrics and social context features. More recent work combined both audio and lyrics-based feature for mood classification [9]. Other cultural data is included in the retrieval process e.g. in the form of textual artist or album reviews [10].

A multi-modal approach to query music, text, and images with a special focus on album covers is presented in [11]. First results for genre classification using the rhyme features used later in this paper are reported in [12]; these results particularly showed that simple lyrics features may well be worthwhile. This approach has further been extended on two bigger test collections, and to combining and comparing the lyrics features with audio features in [13].

III. Audio and Lyrics Features

In this section we describe the audio and lyrics features employed in our experiments. The former comprise Rhythm Patterns, Statistical Spectrum Descriptors, and Rhythm Histograms. The lyrics features are bag-of-words features computed from the terms occurring in the songs, features describing the rhyming structure, features considering the distribution of certain parts-of-speech, and text statistics features.

A. Rhythm Patterns

Rhythm Patterns (RP) are a feature set for handling audio data based on analysis of the spectral audio data and psycho-acoustic transformations [14]. In a pre-processing stage, multiple channels are averaged to one, and the audio is split into segments of six seconds, possibly leaving out lead-in and fade-out segments, and further skipping other segments, e.g. out of the remaining segments every third may be processed.

The feature extraction process for a Rhythm Pattern is composed of two stages. For each segment, the spectrogram of the audio is computed using a Fast Fourier Transform. The window size is set to 1024 samples, applying a Hanning window of 50% overlap. The Bark scale groups frequencies to critical bands according to perceptive pitch regions, is applied to the spectrogram, aggregating it to 24 frequency bands. Then, the spectrogram is transformed into the decibel scale, and further psycho-acoustic transformations are applied. Subsequently, the values are transformed into the unit Sone, on which a doubling on the scale sounds to the human ear like a doubling of the loudness. This results in a psycho-acoustically modified representation reflecting human loudness sensation.

In the second stage, a discrete Fourier transform is applied to the Sonogram, resulting in a (time-invariant) spectrum of loudness amplitude modulation per modulation frequency for each individual critical band. After additional weighting and smoothing steps, a Rhythm Pattern exhibits magnitudes of modulation for 60 modulation frequencies on 24 bands, and has thus 1440 dimensions. Finally, the feature vectors of a songs segments are simply averaged by computing the median.

B. Statistical Spectrum Descriptors

Computing Statistical Spectrum Descriptors (SSD) features [5] relies on the first part of the algorithm for computing RP features, specifically on the Bark-scale representation of the frequency spectrum. From this representation of perceived loudness, seven statistical measures are computed for each of the 24 critical band, to describe fluctuations within them. The statistical measures comprise mean, median, variance, skewness, kurtosis, min- and max-value. A Statistical Spectrum Descriptor is extracted for each segment, and the SSD feature vector of a song is then calculated as the median of its segments. In contrast to the Rhythm Patterns feature set, the dimensionality of the feature space is much lower – SSDs have 168 instead of 1440 dimensions, still at matching performance in terms of genre classification accuracies [5].

C. Rhythm Histograms

The Rhythm Histogram [5] features are a descriptor for the rhythmic characteristics in a song. Contrary to the RP and the SSD, information is not stored per critical band. Rather, the magnitudes of each modulation frequency bin (at the end of the second stage of the RP calculation process) of all 24 critical bands are summed up, to form a histogram of ‘rhythmic energy’ for each of the 60 modulation frequencies. The RH feature vector for a piece of music is calculated as the median of the histograms of each segment. The dimensionality of RH of 60 features is much lower than with the other sets.

D. Bag-Of-Words

Classical bag-of-words indexing at first tokenises all text documents in a collection, most commonly resulting in a set of words representing each document. Let the number of documents in a collection be denoted by N, each single document by d, and a term or token by t. Accordingly, the *term frequency* $tf(t, d)$ is the number of occurrences of term t in document d and the *document frequency* $df(t)$ the number of documents term t appears in. The process of assigning weights to terms according to their importance or significance for a document is called ‘term-weighing’. The weighing we rely on is the most common model, namely the *term frequency times inverse document frequency* [15], computed as $tf \times idf(t, d) = tf(t, d) \cdot ln(N/df(t))$. This results in vectors of weight values for each (lyrics) document d. In this representation it now is possible to compute similarities, as lyrics that contain a similar vocabulary are likely to be semantically related. We did not perform stemming in this setup, earlier experiments showed only negligible differences for stemmed and non-stemmed features [12]; the rationale behind using non-stemmed terms is the occurrence of slang language in some genres.

E. Rhyme Features

Rhyme denotes the consonance or similar sound of two or more syllables or whole words. The reason for considering rhyme as feature is that different genres of music should exhibit different styles of lyrics. We assume the rhyming characteristics of a song to be given by the degree and form of the rhymes used. ‘Hip-Hop’ or ‘Rap’ music, for instance, makes heavy use of rhymes, which (along with a dominant bass) leads to their characteristic sound. To identify such patterns we extract several descriptors from the song lyrics.

Our approach is based on a phoneme transcription of the lyrics. The words ‘sky’ and ‘lie’, for instance, both end with the same phoneme /ai/. The transcription is language dependent; however, our test collection is predominantly composed

TABLE I
RHYME FEATURES FOR LYRICS ANALYSIS

Feature Name	Description
Rhymes-AA	A sequence of two (or more) rhyming lines ('Couplet')
Rhymes-AABB	A block of two rhyming sequences of two lines ('Clerihew')
Rhymes-ABAB	A block of alternating rhymes
Rhymes-ABBA	A sequence of rhymes with a nested sequence ('Enclosing rhyme')
RhymePercent	The percentage of blocks that rhyme
UniqueRhymeWords	The fraction of unique terms used to build the rhymes

of English tracks. After transcribing the lyrics to a phoneme representation, we distinguish two elements of subsequent lines in a song text: *AA* and *AB*. The former represents two rhyming lines, while the latter denotes non-rhyming. Based on these, we extract the rhyme patterns described in Table I. Subsequently, we compute the percentage of rhyming blocks, and define the unique rhyme words as the fraction of unique terms used to build rhymes, describing whether rhymes are frequently formed using the same word pairs. Experimental results indicate that more elaborate patterns based on assonance, semirhymes, or alliterations may well be worth studying.

F. Part-of-Speech Features

Part-of-speech (POS) tagging is a lexical categorisation or grammatical tagging of words. Different POS categories are e.g. nouns, verbs, articles or adjectives. We presume that different genres will differ also in the category of words they are using; thus, we extract several POS descriptors from the lyrics. We count the numbers of: *nouns*, *verbs*, *pronouns*, *relational pronouns* (such as 'that' or 'which'), *prepositions*, *adverbs*, *articles*, *modals*, and *adjectives*. To account for different document lengths, all of these values are normalised by the number of words of the respective lyrics document.

G. Text Statistic Features

Text documents can also be described by simple statistical measures based on word or character frequencies. Measures such as the average length of words or the ratio of unique words in the vocabulary might give an indication of the complexity of the texts, and are expected to vary over different genres. Further, the usage of punctuation marks such as exclamation or question marks may be specific for some genres, and some genres might make increased use of apostrophes when omitting the correct spelling of word endings. The list of extracted features is given in Table II.

All features that simply count character occurrences are normalised by the number of words of the song text to accommodate for different lyrics lengths. 'WordsPerLine' and 'UniqueWordsPerLine' describe the words per line and the unique number of words per line. The 'UniqueWordsRatio' is the ratio of the number of unique words and the total number of words. 'CharsPerWord' denotes the simple average number of characters per word. 'WordsPerMinute' is computed analogously to the well-known beats-per-minute (BPM) value.

TABLE II
OVERVIEW OF TEXT STATISTIC FEATURES

Feature Name	Description
exclamation_mark, colon, single_quote, comma, question_mark, dot, hyphen, semicolon	simple counts of occurrences
d0 - d9	occurrences of digits
WordsPerLine	words / number of lines
UniqueWordsPerLine	unique words / number of lines
UniqueWordsRatio	unique words / words
CharsPerWord	number of chars / number of words
WordsPerMinute	the number of words / length of the song

IV. TEST COLLECTION

The collection we used in the following set of experiments was introduced in [16]. It is a subset of the collection marketed through the content download platform of Verisign Austria (http://www.verisign.at/), and comprises 60,223 of the most popular audio tracks by more than 7,500 artists. The collection contained a number of duplicate songs, which were removed for our experiments. For 41,679 songs, lyrics have been automatically downloaded from portals on the Web. We considered only songs that have lyrics with a certain minimum length, to remove lyrics that are most probably not correctly downloaded.

The tracks are manually assigned by experts to one or more of 34 different genres. 9,977 songs did not receive any ratings at all, and were thus not usable for our genre classification task. Further, we only kept songs that have a rather clear genre categorisation. Thus of those that received more than one voting, we only kept those that have at least two thirds of the votings agreeing on the same genre, removing 12,572 songs. Also, genres with less than 60 songs were not considered.

Finally, after all the removal steps, and thus considering only tracks that have both a clear genre assignment and lyrics in proper quality available, we obtain a collection of 20.109 songs, categorised into 14 genres. Details on the number of songs per genre can be found in Table III. It is noticeable that the different genres vary a lot in size. As such, the smallest class is "Classical", with just 62 songs, or 0.29%. Also, Scores / Soundtrack, Jazz, Blues, Dance / Electronic, Reggae and Comedy comprise less or just about 1% of the whole collection. Contrarily to this, the largest class, Pop, holds 6,156 songs, or 30.6%. Next are two almost equally sized classes, Alternative and Rock, each accounting for almost 3,700 songs or 18.4% of the collection. While this collection clearly is imbalanced towards Pop, Alternative Rock and Rock, accounting for more than two thirds of the collection, it can surely be regarded as a real-world collection. For the experimental results, the class distribution implies a baseline result of the size of the biggest class, thus 30.6%.

V. EXPERIMENTS

In our experiments we compare the performance of audio features and text features using various types of classifiers. We first extracted the audio and lyrics feature sets described

TABLE III
COMPOSITION OF THE TEST COLLECTION

Genre	Artists	Albums	Songs
Pop	1.150	1.730	6.156
Alternative	457	780	3.699
Rock	386	871	3.666
Hip-Hop	308	537	2.234
Country	223	453	1.990
R&B	233	349	1.107
Christian	101	170	490
Comedy	20	44	206
Reggae	30	48	121
Dance / Electronic	48	59	112
Blues	23	39	99
Jazz	38	49	97
Scores / Soundtrack	46	24	70
Classical	21	21	62
Total	3.084	5.074	20.109

in Section III, and then built several dozens of combinations of these different feature sets, both separately within the lyrics modality, as well as combinations of audio and lyrics feature sets. Most combinations are done with the SSD audio features, as this is the best performing audio set. For all experiments, we employed the WEKA machine learning toolkit [1], and unless otherwise noted used the default settings for the classifiers and tests. We used k-Nearest-Neighbour, Naïve Bayes, J48 Decion Trees, Random Forests, and Support Vector Machines. We performed the experiments based on a ten-fold cross-validation. All results given are micro-averaged classification accuracies. Statistical significance testing is performed per column, using a paired t-test with an α value of 0.05.

A. Single Feature Sets

Table IV gives an overview on the classification results. Generally, it can be observed that the results achieved with Naïve Bayes are extremely poor, and are below the above mentioned baseline of the percentage of the largest class, 30.61%, for all but one feature set. SVMs in contrast are performing best on most feature sets, except for those containing text statistics or POS features, where Random Forests are achieving the highest accuracies. In most cases, k-NN achieve better results when the parameter k is increased.

Regarding audio features, shown in the first section of Table IV, the highest accuracies are achieved with SSD features on all classifiers tested; all other feature sets are significantly inferior. Decision Trees on RP and RH features marginally fail to beat the baseline. k-NN seems to improve exceptionally well with a higher value of k – at k of 15, it almost equals the accuracies of the SVMs. Decision Trees and Random Forests perform poorly on RP, just matching the accuracies of RH features. Subsequently, we consider SSD, the highest performing feature set, as the objective we want to improve on in the following experiments on feature combinations.

For lyrics-based rhyme and style features shown in the second part of Table IV, the overall performance is not satisfying. Generally, the text-statistics features are performing best, with the exception of Naïve Bayes, which seems to have some problem with the data, ranking at 2.13%. Rhyme and POS features on SVMs achieve exactly the baseline, where the confusion matrix reveals that simply all instances have been classified into the biggest genre. The part-of-speech features with Naïve Bayes, and the text-statistics on SVM and the Decision Tree manage to marginally outperform the baseline, Decision Trees with statistical significance. Random Forests, however, perform quite well on the text statistics features, achieving almost 40%.

The third section in Table IV gives the results with the bag-of-words features, with different numbers of features selected via frequency thresholding (described in Section III-D). Compared to some of the audio-only features, the results are promising especially with SVMs, yielding the highest accuracy of 51.5%. The SVM classifier is known to perform well with high dimensionalities, which is clearly visible on the BOW features, where the accuracies with SVMs are increasing with the number of features are used. The results with SVMs clearly out-perform the RH features, and with a high dimensionality of 1,500 terms or more achieving even better results than RP; results with 2,000 and more terms show statistically significant improvements. With the other classifiers, the optimal number of features is varying, and not necessarily improving with more dimensions. The best result for k-NN is achieved with k=15 and 1,500 features, yielding 37.44%, clearly below results achieved with audio-only feature sets. With Decision Trees and Random Forests, the results are clearly better than with k-NN, and better than RH or RP features, with Decision Trees even better than the SSD.

Combining the rhyme and style features, slight improvements can be gained in most cases, as seen in the last section of Table IV. Besides Naïve Bayes, the best combination always includes the text statistics and part-of-speech features, and in two out of four cases also the rhyme features. However, the results are still far away from the lyrics features, and not that much better than the baseline. As for POS and text statistics features alone, the best results can be achieved with Random Forests, with around 40% in three of the four combinations.

B. Multi-Modal Feature Set Combinations

Even though the classification results of the lyrics-based features fall short of the SSD features, the objective to improve on, they can be combined to achieve statistically significant improvement over the audio-only results. With the big size of the dataset requiring many computational resources, we focused on combining the lyrics feature sets with SSDs, as they have clearly outperformed the other audio feature sets in the first set of experiments and our previous work.

The second section of Table V shows the results on combining SSD features with the rhyme and style features. It can be noted that each combination performs better than the SSD features alone, except on a few combinations on the k-NN classifier. For all combinations on SVM, most combinations on Decision Trees and Random Forests, as well as some combinations with the k-NN, the improvements are statistically

[1] http://www.cs.waikato.ac.nz/ml/weka/

TABLE IV

CLASSIFICATION ACCURACIES FOR SINGLE AUDIO AND TEXT FEATURE SETS, AS WELL AS RHYME AND STYLE FEATURE COMBINATIONS

Dataset	Dim.	1-NN	3-NN	5-NN	10-NN	15-NN	NB	DT	RF	SVM
RP	1440	38.69	39.41	43.56	45.60	46.13	17.25	30.35	37.72	49.47
RH	60	34.93	35.12	37.94	39.96	40.58	20.50	30.32	37.81	40.75
SSD	168	**46.26**	**46.87**	**49.55**	**51.61**	**52.16**	**26.45**	**39.42**	**48.69**	**55.77**
Rhyme	6	25.29	24.92	26.83	28.00	28.58	22.75	28.13	27.80	30.61
POS	9	27.96	26.61	30.10	31.88	32.78	**32.33**	27.81	32.74	30.61
TextStat	23	**29.40**	**28.32**	**31.56**	**33.27**	**33.57**	2.13	**33.62**	**39.14**	**32.39**
BOW_{20}	20	27.53	25.97	28.11	29.65	30.34	17.83	28.60	n/a	33.11
BOW_{50}	50	30.15	28.90	31.39	33.12	33.52	16.03	31.39	37.50	36.47
BOW_{200}	200	31.36	29.95	32.19	32.88	32.80	19.17	34.12	40.59	43.18
BOW_{399}	399	32.56	30.13	31.29	31.76	31.93	23.58	36.16	41.67	46.30
BOW_{798}	798	**33.20**	29.93	30.79	31.21	31.44	**23.79**	38.75	**42.21**	48.85
BOW_{896}	896	33.06	30.00	30.89	31.52	31.76	23.15	39.02	41.90	49.23
BOW_{997}	997	27.06	22.95	23.68	24.16	28.23	22.30	38.72	41.99	49.42
BOW_{1500}	1500	32.37	31.71	34.39	**36.44**	**37.44**	17.78	39.24	42.04	50.16
BOW_{2000}	2000	32.61	31.83	34.48	35.95	36.90	15.08	40.07	41.55	50.87
BOW_{2232}	2232	32.68	31.68	**34.51**	35.92	36.94	14.29	39.92	41.63	50.92
BOW_{2988}	2988	32.69	31.94	34.12	35.27	35.86	12.98	**41.13**	41.33	51.01
BOW_{3963}	3963	32.90	**32.08**	33.64	34.07	34.17	12.16	41.10	n/a	**51.50**
POS+Rhyme	15	27.91	26.87	29.31	31.08	32.35	**29.00**	27.97	34.33	30.59
POS+TextStat	32	**30.67**	29.66	**32.51**	**34.09**	**35.07**	3.48	**34.19**	**40.42**	35.22
Rhyme+TextStat	29	27.82	26.46	29.30	31.26	32.37	2.33	33.94	39.87	32.54
POS+Rhyme+TextStat	38	30.13	**29.68**	32.31	33.56	34.21	3.82	33.73	40.05	**36.09**

significant. The best result is achieved with SVMs when combining SSD with all of rhyme, part-of-speech and text statistics features. This combination achieves 58.09%, an improvement of 2.3 percent points over the baseline, with only a minor increase in the dimensionality. To offer a comparison on the combination performance, the third section in Table V shows results of combining SSD with RP, RH, and both of them. Only one combination on one classifier leads to significant improvements, while on some classifiers, some combinations even yield significant degradation compared to SSD features only. Also, the highest accuracy for each classifier is topped by several combinations with the rhyme and style features.

Combining SSD with the bag-of-words features, as seen in the third part of Table V, also leads to statistically significant improvements on Decision Trees, Random Forests, and SVMs, for the latter already with only 10 terms used. The best result is achieved on SVM when using around 1500 keyword dimensions with 60.71% classification accuracy, which is statistically significantly better than the SSD combination with the rhyme and style features. It can be noted that generally, k-NN do not profit from adding BOW features. On many combinations, especially with a smaller k, significant degradations can be observed. Only for a few combinations, with around 1,000 to 1,300 terms and higher k values, slight accuracy increased can be gained.

The last two parts of Table V finally present the results of combining all SSD, rhyme and style and bag-of-words features. One of these combinations also achieves the best result in this experiment series, namely SVMs on the last combination presented in the table, with 61.12%.

VI. CONCLUSION

We presented an evaluation on multi-modal features for automatic musical genre classification. Besides features based on the audio signal, we used a set of features derived from song lyrics as an additional, partly orthogonal dimension. Next to measuring the performance of single features sets, we in detail studied the power of combining audio and lyrics features.

Our main contribution in this paper is the large-scale of the evaluation of these features and their combination, on a database of over 20.000 songs. We showed that similar effects as for the smaller, carefully assembled databases of 600 and 3,100 songs presented in earlier work hold true as well for a larger database. Further, the database is taken from a real-world scenario, exhibiting potentially more challenging characteristics, such as having an imbalanced genre distribution.

One interesting observation is that the bag-of-words features alone already achieve very good results, even outperforming RP features, and not being far off the SSD. This, and the improved classification performance achieved on the combination of lyrics and audio feature sets, are promising results for future work in this area. Increased performance gains might be achieved by combining the different feature sets in a more sophisticated approach, e.g. by applying weighting schemes or ensemble classifiers.

REFERENCES

[1] J. Downie, *Annual Review of Information Science and Technology*. Medford, NJ: Information Today, 2003, vol. 37, ch. Music Information Retrieval, pp. 295–340.

[2] N. Orio, "Music retrieval: A tutorial and review," *Foundations and Trends in Information Retrieval*, vol. 1, no. 1, pp. 1–90, September 2006.

[3] J. Foote, "An overview of audio information retrieval," *Multimedia Systems*, vol. 7, no. 1, pp. 2–10, 1999.

[4] G. Tzanetakis and P. Cook, "Marsyas: A framework for audio analysis," *Organized Sound*, vol. 4, no. 30, pp. 169–175, 2000.

[5] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, London, UK, September 11-15 2005, pp. 34–41.

TABLE V

CLASSIFICATION ACCURACIES AND RESULTS OF SIGNIFICANCE TESTING FOR COMBINED AUDIO AND TEXT FEATURE SETS. STATISTICALLY SIGNIFICANT IMPROVEMENT OR DEGRADATION OVER DATASETS (COLUMN-WISE) IS INDICATED BY $(+)$ OR $(-)$, RESPECTIVELY

Dataset	Dim.	1-NN	3-NN	5-NN	10-NN	15-NN	DT	RF	SVM
SSD (test base)	168	46.26	46.87	49.55	51.61	52.16	39.42	48.69	55.77
SSD + POS	177	46.04	47.06	50.06	52.39	52.48	39.78	49.96 +	56.52 +
SSD + TextStat	191	46.94	47.70	50.58 +	52.22 +	52.95	42.46 +	51.13 +	57.35 +
SSD + Rhyme + TextStat	197	46.56	47.06	50.57	52.21	52.78	41.74 +	50.94 +	57.42 +
SSD + POS + TextStat	200	46.97	47.84 +	50.71 +	52.65	53.24 +	42.40 +	50.68 +	57.92 +
SSD + POS + Rhyme + TextStat	206	46.59	47.79 +	50.72 +	52.78 +	53.21 +	41.60	50.82 +	58.09 +
SSD + RP	1608	44.39 -	45.37 -	48.60	50.75	51.53	35.69 -	43.63 -	55.77
SSD + RH	228	46.75	47.19	50.24	52.19	52.77	39.12	48.96	57.42 +
SSD + RP + RH	1668	44.38 -	45.62 -	48.69	50.59	51.45	36.51 -	43.59 -	55.60
BOW_{10} + SSD	178	39.87 -	40.20 -	43.90 -	46.13 -	46.30 -	39.45	48.78	56.32 +
BOW_{20} + SSD	188	40.11 -	40.83 -	44.95 -	46.95 -	48.00 -	39.38	49.14	57.12 +
BOW_{50} + SSD	218	41.52 -	43.42 -	46.68 -	48.71 -	49.70 -	39.54	49.36	58.46 +
BOW_{75} + SSD	243	41.86 -	43.45 -	46.21 -	49.01 -	49.56 -	39.78	48.74	58.84 +
BOW_{150} + SSD	318	41.98 -	43.44 -	46.64 -	48.55 -	49.80 -	39.10	49.16	59.17 +
BOW_{300} + SSD	468	42.05 -	42.77 -	45.22 -	48.10 -	50.00 -	40.82 +	48.62	59.98 +
BOW_{798} + SSD	966	39.50 -	42.46 -	47.33 -	50.52	51.64	41.72 +	48.32	60.49 +
BOW_{997} + SSD	1165	42.33 -	44.69 -	49.08	51.52	52.47	42.01 +	48.05	60.69 +
BOW_{1248} + SSD	1416	43.91 -	45.92	49.58	51.57	52.15	42.26 +	48.37	60.51 +
BOW_{1500} + SSD	1668	44.15 -	45.61 -	49.29	50.80	51.59	42.22 +	47.70 -	60.71 +
BOW_{2232} + SSD	2400	43.13 -	45.10 -	48.18 -	49.68 -	49.55 -	42.84 +	47.01 -	60.70 +
BOW_{2988} + SSD	3156	42.02 -	43.87 -	46.96 -	47.91 -	47.99 -	43.52 +	47.33 -	60.41 +
BOW_{3963} + SSD	4131	41.45 -	43.01 -	45.17 -	46.79 -	46.73 -	43.36 +	n/a	60.67 +
BOW_{10} + SSD + TextStat	201	40.46 -	41.04 -	44.28 -	46.29 -	46.73 -	42.09 +	50.70 +	57.59 +
BOW_{20} + SSD + TextStat	211	40.61 -	41.54 -	44.91 -	47.63 -	48.52 -	42.75 +	50.69 +	58.14 +
BOW_{50} + SSD + TextStat	241	42.40 -	43.89 -	47.12 -	49.04 -	49.90 -	42.54 +	50.07	59.16 +
BOW_{150} + SSD + TextStat	341	43.00 -	44.19 -	47.22 -	49.43 -	50.47 -	41.26 +	50.35 +	59.95 +
BOW_{399} + SSD + TextStat	590	41.99 -	42.55 -	45.11 -	49.16 -	51.00	41.01 +	50.25 +	60.55 +
BOW_{799} + SSD + TextStat	989	40.80 -	44.11 -	48.64	51.21	52.13	41.51	49.34	60.68 +
BOW_{997} + SSD + TextStat	1188	45.26	46.76	49.93	51.86	52.90	41.63 +	49.10	60.54 +
BOW_{1248} + SSD + TextStat	1439	44.30 -	46.43	49.83	52.26	52.80	42.10 +	48.91	60.66 +
BOW_{1500} + SSD + TextStat	1691	44.23 -	46.22	49.54	51.79	52.68	42.05 +	48.37	60.87 +
BOW_{2232} + SSD + TextStat	2423	43.73 -	45.83	48.80	50.34	50.73 -	42.93 +	47.86	60.76 +
BOW_{2988} + SSD + TextStat	3179	42.38 -	44.49 -	47.35 -	48.47 -	48.73 -	43.32 +	47.40	60.63 +
BOW_{3963} + SSD + TextStat	4154	41.81 -	43.06 -	46.00 -	47.40 -	47.23 -	43.54 +	n/a	61.05 +
BOW_{10} + SSD + TextStat + POS + Rhyme	216	41.05 -	41.91 -	45.11 -	47.00 -	47.40 -	41.60	50.65 +	58.39 +
BOW_{20} + SSD + TextStat + POS + Rhyme	226	41.71 -	42.41 -	45.95 -	47.80 -	48.79 -	42.36 +	50.14 +	58.90 +
BOW_{50} + SSD + TextStat + POS + Rhyme	256	43.26 -	44.63 -	47.77 -	49.83 -	50.66 -	42.24 +	50.51 +	59.62 +
BOW_{150} + SSD + TextStat + POS + Rhyme	356	44.64 -	45.49 -	48.61	50.69 -	51.52 -	40.78	50.29 +	60.07 +
BOW_{399} + SSD + TextStat + POS + Rhyme	605	43.96 -	44.45 -	47.90 -	51.09	52.96	41.09	50.23 +	60.72 +
BOW_{798} + SSD + TextStat + POS + Rhyme	1004	43.22 -	45.80	50.28	52.57	53.56 +	41.58 +	48.83	60.67 +
BOW_{997} + SSD + TextStat + POS + Rhyme	1203	45.68	47.70 +	50.98 +	53.39 +	53.78 +	41.45 +	48.91	60.86 +
BOW_{1248} + SSD + TextStat + POS + Rhyme	1454	44.88 -	47.40	50.96 +	53.05 +	53.76 +	42.38 +	48.71	61.06 +
BOW_{2232} + SSD + TextStat + POS + Rhyme	2438	43.82 -	46.44	49.38	51.68	52.13	43.02 +	47.29 -	60.92 +
BOW_{2988} + SSD + TextStat + POS + Rhyme	3194	42.85 -	44.85 -	48.27 -	49.59 -	49.98 -	43.23 +	47.58	60.79 +
BOW_{3963} + SSD + TextStat + POS + Rhyme	4169	41.87 -	43.72 -	46.66 -	48.13 -	48.34 -	43.48 +	n/a	61.12

[6] J. P. G. Mahedero, Á. Martínez, P. Cano, M. Koppenberger, and F. Gouyon, "Natural language processing of lyrics," in *Proceedings of the ACM 13th International Conference on Multimedia (MM'05)*, New York, NY, USA, 2005, pp. 475–478.

[7] B. Logan, A. Kositsky, and P. Moreno, "Semantic analysis of song lyrics," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'04)*, Taipei, Taiwan, June 27-30 2004, pp. 827–830.

[8] D. Yang and W. Lee, "Disambiguating music emotion using software agents," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, October 2004.

[9] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," San Diego, CA, USA, December 11–13 2008, pp. 688–693.

[10] S. Baumann, T. Pohle, and S. Vembu, "Towards a socio-cultural compatibility of mir systems." in *Proceedings of the 5th International Conference of Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, October 10-14 2004, pp. 460–465.

[11] E. Brochu, N. de Freitas, and K. Bao, "The sound of an album cover: Probabilistic multimedia and IR," in *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, C. M. Bishop and B. J. Frey, Eds., Key West, FL, USA, January 3-6 2003.

[12] R. Mayer, R. Neumayer, and A. Rauber, "Rhyme and style features for musical genre classification by song lyrics," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*, September 14-18 2008, pp. 337–342.

[13] ——, "Combination of audio and lyrics features for genre classification in digital audio collections," in *Proceedings of the ACM Multimedia 2008*, Vancouver, BC, Canada, October 27-31 2008, pp. 159–168.

[14] A. Rauber, E. Pampalk, and D. Merkl, "Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by musical styles," in *Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR'02)*, Paris, France, October 13-17 2002, pp. 71–80.

[15] G. Salton, *Automatic text processing – The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[16] F. Kleedorfer, P. Knees, and T. Pohle, "Oh oh oh whoah! towards automatic topic detection in song lyrics," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, Philadelphia, PA, USA, September 14 – 18 2008, pp. 287 – 292.